Cardiovascular Disease Prediction using Machine Learning and Deep Learning

Thirupati Sai Eswar Reddy[1], Satwik Reddy Sripathi[1], Dhanush Akula[1], Suja Palaniswamy1*, Subramani R[2]
[1]*Department of Computer Science and Engineering, Amrita School of Computing, Bengaluru, Amrita Vishwa Vidyapeetham, India.*
[2]*Department of Mathematics, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India.*
saieswarreddytirupati01@gmail.com, satwikreddysripathi@gmail.com, dhanushakula18@gmail.com, p_suja@blr.amrita.edu, r_subramani@blr.amrita.edu

***Abstract* — One of the leading causes of death is CVD(Cardiovascular disease) which is communally referred to as heart disease. CVD is the summation of disorders that affect the heart's ability to function. Every year, over 18 million people worldwide die as a result of heart disease. One in three deaths from cardiovascular disease in them is preventable, and heart attacks can be predicted months in advance by assessing the patient's risk factors. Obesity, Blood Pressure, Cholesterol, and Glucose levels are some of the risk factors. The aim is to predict CVD based on risk factors of patients, who are affected by their habits and patients' basic health information, using the Cardiovascular Disease Dataset from Kaggle using Machine Learning and Deep Learning Algorithms. Smoking and Alcohol consumption are some practices that will maximize the possibility of getting cardiovascular disorders, and doing workouts will reduce the risk of getting cardiovascular disease. In our work, we have implemented four models for predicting CVD namely Logistic Regression, Naive Bayes, Deep Learning Model and Random Forest. The DL model, with an accuracy of 73.78 %, outperformed the other three models.**

Keywords— Machine Learning, Cardiovascular Disease, Random Forest, Logistic Regression, Naïve Bayes, Multi-layer perceptron

I. INTRODUCTION

The heart is an essential and important organ in the human body. It is the main part of the circulatory system. The fluid assists in the transport of oxygen, which is required for the body's proper operation. Each year, the heart pumps enough blood to fill a modern tanker. Approximately four million times every year it beats [1]. Cardiovascular disorders are the most common cause of mortality worldwide. CVDs cost the lives of 17.9 million people worldwide in 2022, contributing to 32% of all deaths. Heart attacks and strokes are reported for more than 80% of CVD deaths, and 33% of these deaths occur in adults under the age of 70. Physical inactivity, Unhealthy eating, cigarette smoking, and alcohol intake are some of the key behavioral risk factors that led to heart disease and stroke. It's crucial to diagnose CVD as early as possible so that behavioral therapy and medication can begin. in. Rise in blood glucose, blood pressure, blood lipids, and overweight are experienced by individuals as a result of behavioral risk factors. According to the World Health Organization (WHO), heart disease affects both genders equally and 17.9 million humans died of CVD within a year. Heart attacks and Stroke account for over 85% of all global mortality [2].

One-fifth of deaths have been linked to cardiac and occur when a trigger interacts with an arrhythmic substrate [3]. Along with cardiovascular disorders, there are some issues. There can be a case whereby the arteries harden and become inflexible and thicker [19]. This is called arteriosclerosis. We can also have atherosclerosis which is the narrowing of the arteries that reduce blood flow. A blood clot or obstruction in the heart's blood flow is the most common cause of heart attacks. On September 22, 2016, the World Health Organization (WHO) started the Global Hearts campaign to emphasize the significance of preventing fatalities from cardiovascular illnesses. es. High cholesterol, smoking, High blood pressure, Physical inactivity, unhealthy diet, obesity, and poorly managed diabetes are all the risk factors for heart disease. A medical history, a stethoscope, an ultrasound, and an ECG are commonly used to diagnose cardiac disease.

Data mining is a technique for identifying meaningful patterns in a vast amount of data. Understanding and producing discoveries from data is the most crucial task in data mining. It has recently attracted the attention of the majority of academics [4]. Strong analysis tools are anticipated to be required as the volume of the health report collected through the EHR systems grows. With such a large volume of data, health care providers are now using data mining to improve their organization's efficiency. Data mining has notably assisted the healthcare industry in reducing costs by enhancing efficiencies. It has aided in improving the patient's quality of life, consequently preserving patients' lives. Predictive medicine, fraud detection, customer relationship management, healthcare management, and estimating the success of specific therapies have all benefited from data mining. Data mining can be applied to different applications relating to health. The following categories have been used to categorize these applications.

By making clinical decisions, Clinicians typically examine patients to diagnose their diseases. There is a potential that the diagnosis will be inaccurate because this is an experimental approach. For most diagnoses, data mining provides a second opinion to the practitioner. This helps clinicians make more accurate forecasts by ensuring that the condition is not underestimated during diagnosis. The state of the public health Epidemiologists, for example, study disease prevalence and are usually interested in finding the trends, patterns, and causes of the disease throughout the population. They take into account early life, lifestyle, and social-demographic aspects [5].

Research works conducted on the UCI heart disease dataset so far only contains the lab results of patients, this work dives further. This research work looks at using patient's basic health information along with their habits

to determine the chances of them having a CVD. In this work we have used cardiovascular disease dataset and developed a Deep learning model to predict CVD and compared the results with other machine learning models.

II. Related Work

CVD-related mortality has decreased in several wealthy countries, but have grown significantly in countries with low and medium incomes. Around 80% of the global burden is borne by these countries. Asian Indians have a mortality rate associated with coronary artery disease that is 20–50 percent greater than the general population [6]. Artificial intelligence (AI)-based applications have found extensive use in a variety of scientific, technological, and medical domains. Since the 1960s, researchers have been investigating the use of machines with increased computational capacity in clinical medicine and diagnostics. There are many various uses in the Ai based systems for cardiovascular care such as cardiovascular risk prediction, cardiovascular imaging, and novel treatment targets. [7]. Alcohol use of three or more drinks per day, as well as cigarette smoking, have unfavorable consequences on specific forms of CVD that are comparable and possibly cumulative [21]. Increased blood pressure and lipid levels in the blood, as well as increased risks of stroke and congestive heart failure, are examples of these negative effects. There is tentative evidence that the two function synergistically or that the results are worse when smoking and drinking happen at the same time than would be assumed from their separate effects [8]. Diabetes patients are majorly face to develop heart failure, while those with heart failure are more likely to get diabetes. Furthermore, antidiabetic medications increase the possibility of death and heart failure hospitalization in individuals with and without pre-existing heart failure. [9].

Physical activities will be effective to lower blood pressure, improving fibrinolytic capacity, decreasing blood coagulation, and helping in vascular remodeling. Smoking raises the risk of atherosclerosis and interferes with the cardiopulmonary system's proper function. Quitting smoking is highly recommended because it can be a risk factor that is avoidable for cardiovascular and respiratory-related diseases [10]. High BP is a serious possibility factor for CVD and the leading cause of mortality. High blood pressure caused around 54% of strokes and 47% of coronary heart disease [11]. Hypercholesterolemia is a preventable cause for coronary heart disease (CHD). High blood cholesterol along with smoking and hypertension have been sighted as a primary modifiable risk factor for CHD. Both the Lipid Research Clinics Coronary Primary Prevention Trial and the National Heart, Lung, and Blood Institute Type II Coronary Intervention Study recently revealed that reducing cholesterol levels lowers the incidence of CHD and delays the course of atherosclerosis. Diet therapy alone can reduce blood cholesterol levels by 10% to 15%, whereas a combination of diet and medications can reduce levels by 30% or more[12].

Heart and blood vessel disorders caused by cardiovascular disease (CVD) often lead to death or paralysis. Therefore, early and automatic detection of cardiovascular disease could prevent many lives from being lost. Some research has been done to reach this goal, but performance and reliability can still be improved. This paper is another step in that direction. In this study [13], using two of his highly reliable machine learning methods, multi-layer perceptron (MLP) and K-nearest neighbors (K-NN), he used data from the University of California, Irvine repository to of CVDs were detected [17]. To improve model performance, outliers and features with zero values should be removed. Experimental data show that the MLP model outperforms the K-NN model in terms of detection accuracy (82.47%) and area under the curve (86.41%). As a result, the suggested MLP model for automatic CVD detection was suggested. Other diseases can be detected using the suggested methods. Additionally, other common data sets can be used to evaluate the performance of the proposed model. A precise heart disease diagnosis can lower the likelihood of developing major health issues, whereas an incorrect diagnosis can be fatal. In this work, a variety of machine learning approaches, including deep learning, are utilized to compare the results and analyses of the UCI Machine Learning Heart Disease dataset [20]. There are 14 important attributes in the dataset that will be used in the study. Several encouraging results are obtained and validated using the accuracy and confusion matrix. The dataset is normalized for better results and some minor features are managed using isolation forests. The potential integration of this study with other multimedia technology, such as mobile devices, is also noted. The deep learning approach resulted in 94.2% accuracy in this paper [14].

The present advancement in artificial intelligence is crucial in aiding medical professionals in their diagnosis of various disorders. In the current work [15], a hybrid framework for the use of medical speech records in the identification of cardiac problems is proposed. It is suggested to use a framework with four layers: "Segmentation," "Features Extraction," "Learning and Optimization," and "Export and Statistics.". An innovative segmentation method based on the segmentation of varying durations and directions is proposed in the first layer. 11 datasets with 14,416 numerical features are produced using the suggested method [18]. The task of feature extraction falls under the second layer. The obtained datasets are used to extract numerical and graphical features. In the third layer, graphical features are transmitted to 8 distinct Convolutional Neural Networks with the help of transfer learning to find the best configurations, while the numeric features are passed to 5 different Machine Learning methods. The hyperparameters of the ML and CNN configurations are optimized using Grid Search and Aquila Optimizer (AO), respectively. The validation of the suggested hybrid framework result is done using various performance measures in the final layer. The two best reported scores are (2) 99.17% accuracy using CNN, and (1) 100% accuracy utilizing ML techniques, including the Extra Tree Classifier and the Random Forest Classifier (RFC).

The rest of the paper is organized as follows:
Section II briefs about the related and state-of-the art results which motivated this research work, Section III details about the proposed methodology and the flow of the work and in Section-IV, final results are detailed.

III. PROPOSED METHOD

The workflow and the methodology of the proposed work in identifying the harmful cardiovascular disease is detailed in brief in this section.

A. Network Design:

The flow of the entire network of all models used for this research is described by network design.

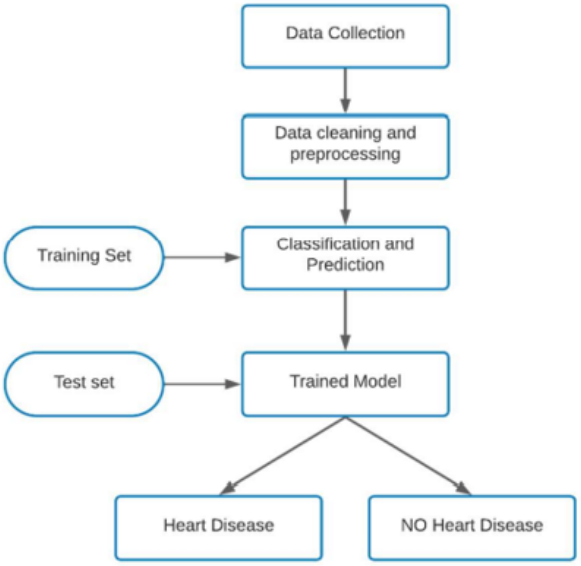

Fig. 1. System Workflow

B. CardioVascular Disease Dataset:

This research uses the cardiovascular disease dataset from Kaggle. Three kinds of input features are present : Objective: factual information, Examination: the outcome of the medical examination, Subjective: information taken from patients.It contains 70000 samples with 13 attributes. Table-1 display's the cardiovascular Disease Dataset's detailed description.

Table 1. Attributes of the Cardiovascular Disease Dataset

Age	Objective Feature	Age	int(days)
Height	Objective Feature	Height	int(cm)
Weight	Objective Feature	Weight	float(kg)
Gender	Objective Feature	Gender	Categorical Code
Systolic BP	Examination Feature	ap_hi	int
Diastolic BP	Examination Feature	ap_lo	int
Cholesterol	Examination Feature	Cholesterol	1:Normal 2:Above Normal 3:Well Above Normal
Glucose	Examination Feature	Gluc	1:Normal 2:Above Normal 3:Well Above Normal
Smoking	Subjective Feature	Smoke	Binary
Alcohol Intake	Subjective Feature	Alco	Binary
Physical Activity	Subjective Feature	Active	Binary
Presence or Absence of CardioVascular Disease.	Target Variable	Cardio	Binary

C. Data Processing:

Cleaning data is another term for data preprocessing. It is one of the most crucial processes in getting the most out of the dataset. This is a method for reducing data inconsistencies. In this step, we will remove the duplicate values, Then find BMI values using weight and height with simple math and add it as a feature.[16] Next, converting age which is in days to years, removing age, height, weight features from the dataset.

D. Performance Evaluation Metric:

The prediction accuracy of a classification algorithm is assessed using a classification report. How accurate were your predictions, and how many were incorrect? The metrics of a categorization report are predicted using True Positives, True Negatives, False Negatives and False Positives. The classification metrics are precision, recall, F1-score, Accuracy

E. Structure Learning For Multi-Layer Perceptron Network:

After cleaning our data to make it modellable, we utilize TensorFlow and Kera's in Python to build an MLP Network. There are 5 layers present in our model: starting with an input layer, continuing with 2x (hidden layer and dropout layer). In the input layer, we have 10 nodes as from the 11 attributes we have, one is the output attribute. As this model is created for binary classification one node is present in the output layer.

The total number of neurons inside the hidden layers must be determined as part of the conventional neural community architecture. Other than the fact that such layers no longer interact with the outside world at the same time, they have a significant influence on the end result. The number of neurons as well as the number of hidden layers in each of those hidden layers must be properly estimated.

A small Gaussian random number is used to initialize the weights. The activation function of the Rectifier is employed. In order to make predictions, the output layer has a single neuron. It employs the sigmoid activation function to generate a probability output ranging from 0 to 1 that could be easily and automatically translated into precise class values.

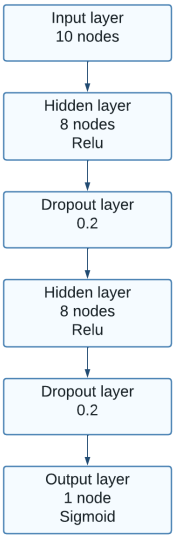

Fig. 2. Deep Learning Model Architecture

As shown in the Fig.2 1st hidden layer contains 8 nodes with relu activation function followed by that we have Dropout layer with 0.2 dropout rate the 1st and 2nd were repeated again then finally we have output layer with 1 node and sigmoid activation function.

Finally, during training, we employ the logarithmic loss function (binary cross entropy), which is the preferred loss function for binary classification problems. When the model is trained, accuracy measurements will be collected and the model will employ the efficient Adam optimization technique for gradient descent.

IV. Results and Discussion

With the work described above, we were successfully able to achieve the following results. Table 2 represents the results generated after applying the mentioned algorithms. Fig 3,4 represents model accuracy and loss.

Table 2. Comparison of the Models

Model	Precision	recall	F1-score	Accuracy
Naive Bayes	0.72	0.71	0.71	70.88
Logistic Regression	0.73	0.73	0.73	73.01
Random Forest	0.71	0.71	0.71	70.56
Deep Learning Model	0.74	0.74	0.74	73.78

.

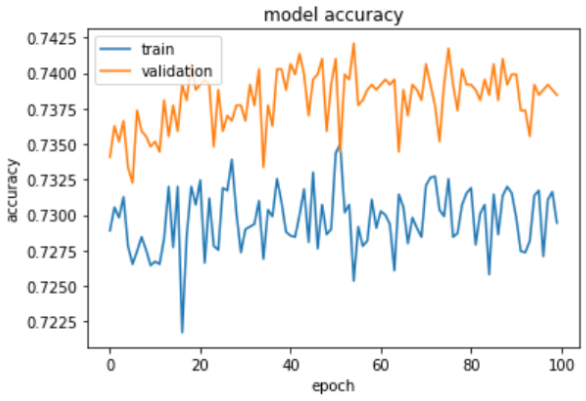

Fig. 3. Performance of the DL Model

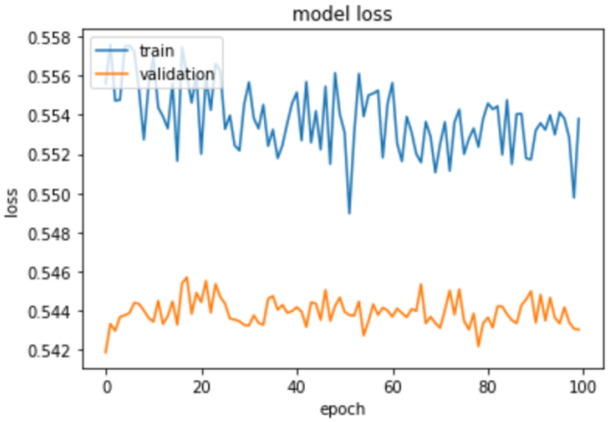

Fig. 4.Loss of the DL Model

V. Conclusion and Future Scope

In this paper, a novel approach of detecting cardiovascular related diseases based on the habits and history of health of an individual are detected. Based on all this information, as part of this research, a set of algorithms are developed to predict a patient's CVD based on basic health information and habits that affect the patient risk factors that lead to CVD. As a part of the research, Kaggle CVD dataset with 70000 records and 11 features + Target. Logistic Regression, Naive Bayes, Random Forest, and the DL Model are the algorithms used. The DL model outperformed the other three models with an accuracy of 73.78%.

References

[1]. S. B. Meaghan George, S. heart Bleiberg, N. Alawa and D. Sanghavi. Case study: Delivery and payment reform in congestive failure at two large academic centers. Elsevier, vol. 2, pp. 107-112, 2014

[2]. www.who.int/health-topics/cardiovascular-diseases

[3]. van der Bijl P, Delgado V, Bax JJ. Imaging for sudden cardiac death risk stratification: Current perspective and future directions. Prog Cardiovasc Dis. 2019 May-Jun;62(3):205-211. doi: 10.1016/j.pcad.2019.04.005. Epub 2019 May 2. PMID: 31054859.

[4]. Wu Y, He Z, Lin H, Zheng Y, Zhang J, Xu D. A Fast Projection-Based Algorithm for Clustering Big Data. Interdiscip Sci. 2019 Sep;11(3):360-366. doi: 10.1007/s12539-018-0294-3. Epub 2018 Jun 7. PMID: 29882026.

[5]. Mohammad Hossein Tekieh and Bijan Raahemi. 2015. Importance of Data Mining in Healthcare: A Survey. In Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015 (ASONAM '15). Association for Computing Machinery, New York, NY, USA, 1057–1062.

[6]. Sreeniwas Kumar A, Sinha N. Cardiovascular disease in India: A 360 degree overview. *Med J Armed Forces India*. 2020;76(1):1-3. doi:10.1016/j.mjafi.2019.12.005

[7]. Mathur P, Srivastava S, Xu X, Mehta JL. Artificial Intelligence, Machine Learning, and Cardiovascular Disease. *Clin Med Insights Cardiol*. 2020;14:1179546820927404. Published 2020 Sep 9. doi:10.1177/1179546820927404

[8]. Mukamal KJ. The effects of smoking and drinking on cardiovascular disease and risk factors. *Alcohol Res Health*. 2006;29(3):199-202.

[9]. Rosano GM, Vitale C, Seferovic P. Heart Failure in Patients with Diabetes Mellitus. *Card Fail Rev*. 2017;3(1):52-55. doi:10.15420/cfr.2016:20:2

[10]. Buttar HS, Li T, Ravi N. Prevention of cardiovascular diseases: Role of exercise, dietary interventions, obesity and smoking cessation. *Exp Clin Cardiol*. 2005;10(4):229-249.

[11]. Wu CY, Hu HY, Chou YJ, Huang N, Chou YC, Li CP. High Blood Pressure and All-Cause and Cardiovascular Disease Mortalities in Community-Dwelling Older Adults. *Medicine (Baltimore)*. 2015;94(47):e2160. doi:10.1097/MD.0000000000002160

[12]. Clark LT. Cholesterol and heart disease: current concepts in pathogenesis and treatment. *J Natl Med Assoc*. 1986;78(8):743-751.

[13]. Pal M, Parija S, Panda G, Dhama K, Mohapatra RK. Risk prediction of cardiovascular disease using machine learning classifiers. Open Med (Wars). 2022 Jun 17;17(1):1100-1113. doi: 10.1515/med-2022-0508. PMID: 35799599; PMCID: PMC9206502.

[14]. Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, Parneet Singh, and Ahmed A. Abd El-Latif. 2021. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Intell. Neuroscience 2021 (2021). https://doi.org/10.1155/2021/8387680

[15]. Balaha, H.M., Shaban, A.O., El-Gendy, E.M. *et al.* A multi-variate heart disease optimization and recognition framework. *Neural Comput & Applic* **34**, 15907–15944 (2022). https://doi.org/10.1007/s00521-022-07241-1

[16]. Suja P., Shikha Tripathi, "Emotion recognition from Facial Expressions using 4D Videos and Analysis on Feature-Classifier Combination," Int'l J. Intelligent Engineering and Systems, vol.10, no.2, pp.30-39, The Intelligent Networks and Systems Society (INASS), Japan, April 2017.

[17]. A. J. B and S. Palaniswamy, "Comparison of Conventional and Automated Machine Learning approaches for Breast Cancer Prediction," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1533-1537, doi: 10.1109/ICIRCA51532.2021.9544863

[18]. K. S. Srikanth, T. K. Ramesh, S. Palaniswamy and R. Srinivasan, "XAI based model evaluation by applying domain knowledge," 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2022, pp. 1-6, doi: 10.1109/CONECCT55679.2022.9865816

[19]. S. S., S., R., and Jisha R. C., "A Real Time Patient Monitoring System for Heart Disease Prediction Using Random Forest Algorithm", in Advances in Signal Processing and Intelligent Recognition Systems, Cham, 2016, vol. 425, pp. 485-500

[20]. R. Prasanna Kumar, "An empirical study on machine learning algorithms for heart disease prediction", IAES International Journal of Artificial Intelligence , 2021\

[21]. Manju Priya Arthanarisamy Ramaswamy, Suja Palaniswamy, "Subject independent emotion recognition using EEG and physiological signals – a comparative study",Applied Computing and Informatics. ISSN: 2634-1964